Applied Psycholinguistics, ***22*** *(2001), 509-540*

Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints

John N. Williams and Peter Möbius

University of Cambridge, UK

Choonkyong Kim

Chonnam National University, South Korea

Two experiments are reported which investigated the processing of English wh- questions by native speakers of English and advanced Chinese, German and Korean learners of English as a second language. Performance was evaluated in relation to parsing strategies and sensitivity to plausibility constraints. In an on-line plausibility judgement task both natives and non-natives behaved in similar ways. All groups postulated a gap at the first position consistent with the grammar, as predicted by the "filler-driven" strategy, and as shown by garden-path, or "filled gap", effects that were induced when the hypothesised gap location turned out not to be correct. All subjects also interpreted the plausibility of the filler-gap dependency, as was shown by a reduction in the garden-path effect when the initial analysis was implausible. However, the native reading profiles showed evidence of a more immediate effect of plausibility than the non-natives, suggesting that they initiated reanalysis earlier when the first analysis was implausible. Experiment 2 showed that only non-natives had difficulty cancelling a plausible gap hypothesis even in an off-line (pencil and paper) task, whereas for natives there was no evidence that the sentences caused difficulty in this situation. The results suggest that natives and non-natives employ similar strategies in immediate on-line processing, and hence are garden-pathed in similar ways, but they differ in their ability to recover from misanalysis.

Studies of second language acquisition have been dominated by investigations of learners' knowledge of language, be it of lexis, morphology, or syntax. In the area of syntax in particular, attention has focused upon whether the adult learner can approach the same level of knowledge possessed by a native speaker as indicated by, for example, intuitions about the acceptability of sentences, and whether the ability to acquire such knowledge is determined by the nature of their first language grammar. Within the Principles and Parameters framework this question is framed

Experiment 1 was performed with the support of the Economic and Social Research Council of Great Britain (grant number N000221731). We thank Ngoni Chipere for providing the software used to run Experiment 1, Kerrie Elston-Güttler for collecting the German data in Experiment 2, Yi'an Wu and Suhua Jiang for collecting the Chinese data for Experiment 2, Ernest Lee and Teresa Parodi for helpful discussion and advice, and two anonymous reviewers for helpful and insightful comments Address correspondence to John N. Williams, jnw12@cam.ac.uk.

in terms of whether the learner can reset parameters to values in the L2 that differ from those in the L1 (Towell & Hawkins, 1994). From this perspective, once it has been demonstrated that a structure is known then it seems to be generally assumed that there are no further interesting questions that can be asked about its acquisition.

However, from a psycholinguistic perspective, knowing a structure, in the sense of accepting sentences that contain it as grammatical, or even successfully interpreting or producing them, constitutes only the starting point for investigations of how the relevant grammatical knowledge is put to use during language processing. In those areas where a consensus is emerging concerning the processing routines and strategies that a native speaker of a language employs it is possible to ask whether those same routines and strategies can be identified in learners of a language who appear to know the relevant structures. One can also ask whether there is any difference in the processing routines that learners appear to use depending upon the nature of their first language; in particular, how similar that language is to the target language in the relevant respects.

The present study investigated these issues in the context of wh- constructions, an area of syntax where there is a relatively high degree of consensus regarding the nature of native language parsing strategies. Most of this research has followed the generative approach, as in, for example, Government Binding theory (Chomsky, 1981), and analysed wh- questions in terms of wh-traces [t]. For example, in (1) the trace [t] is posited as an (invisible) surface marker of the original site of the filler in the d-structure of the sentence.

(1) Bill wonders [what]$_i$ the manager wants the assistant to put [t_i] in the sales next week.

Such structures raise interesting questions concerning the way in which the grammar is implemented in the human parsing mechanism. Since traces are invisible in the surface form of the sentence they must be inferred, so one may wonder what strategies are used to establish where a trace is located. Two aspects of parsing English wh-questions will be examined: first, the strategies that are used for identifying gap sites, and second, the influence of plausibility constraints. These issues will first be discussed in relation to research on native speakers of English.

Processing wh-questions by native speakers of English

Identifying the gap site

The first major processing problem posed by even simple wh- questions is that of the identification of the gap site. Since there are no overt markers in the surface form to indicate that a gap exists, its presence must be inferred. The first, and simplest strategy, is to wait until a sequence of words is encountered which would not be syntactically well-formed unless a gap were postulated and the filler inserted at that position. For example, in (1) the reader might read *.... the manager wants the assistant to put in* and realise that the sequence *put in* can not be parsed unless the filler *what* is inserted after *put*. This was the solution to gap identification adopted by the Augmented Transition Network parser of Wanner & Maratsos (1978), and was referred to by Fodor (1978) as the "Gap as Last Resort " strategy. An alternative, less obvious, solution is that the reader, having identified a wh-filler (e.g. *what* in sentence (1) then enters a processing mode in which they actively predict where there might be a gap. This strategy allows a gap to be hypothesised before an otherwise unparsable sequence of words is encountered. For example, the reader would process *Bill wonders what the manager wants*, posit a gap after *wants*, and immediately interpret *what* as the direct object of that verb. On encountering *the assistant* this analysis would have to be cancelled, and the reader would have to wait for the next potential

gap position that is licensed by the grammar. Frazier & Clifton (1989) refer to this as the "filler-driven strategy".

According to Frazier & Clifton (1989) skilled native speakers of English adopt the filler-driven strategy. One compelling line of evidence for this hypothesis comes from a study by Stowe (1986) who, using materials similar to (1), showed that it is indeed the case that readers slow down after a potential but not realised gap position, suggesting that they are forced to reanalyse the structure of the sentence at that point. Specifically, Stowe showed that in sentences such as (2) reading times slow down after the verb *bring*.

(2) My brother wanted to know who Ruth will bring us home to at Christmas.

It was argued that this is because the reader assumes that there is a gap after *bring* which is filled by *who*, and that this interpretation has to be revised upon encountering *us*. In other words, the reader slows down at points where a gap might be predicted on structural grounds but turns out not to exist because it is in a sense "filled" by an overt element of surface structure. This specific type of garden-path effect is referred to as the "filled-gap effect".

The Filler Driven strategy is instantiated in many current models of parsing (e.g. Gibson, Hickok, & Shütze, 1994; Pritchett, 1992) and is even compatible with models which deny the psychological reality of traces altogether (Pickering & Barry, 1991; Pickering, 1994). This is because it reflects a general property of sentence processing, at least in native speakers, which Just & Carpenter (1980) referred to as the "immediacy hypothesis". During comprehension, the reader or listener attempts to interpret the input in as incremental a fashion as possible. Notwithstanding the wide agreement on this fundamental principle, there has been much debate on what is meant by "interpret". The existence of garden-path phenomena attests to the immediacy of syntactic analysis, but is semantic interpretation immediate as well, and if so, do the results of semantic analysis (in the form of degree of plausibility) have an immediate influence on sentence processing?

Plausibility constraints

There has been some debate over the extent to which gap processing, specifically the filler-driven strategy, is subject to subcategorization and plausibility constraints. With regard to subcategorization, Stowe, Tanenhaus, & Carlson (1991) provide evidence that there is no filled-gap effect for intransitive-preferred verbs (e.g. after *hurried* in *The nurse wondered which patient the orderly hurried quickly towards*), suggesting that direct object gaps are not routinely posited in such cases. In contrast, the same study showed that when a transitive verb is used, slower reading times are found at the verb when the filler is implausible as its direct object than when it is plausible, even when in both cases there is in fact no gap at that position. For example, the time spent reading *read* is greater in *The teacher wondered which song the students read quietly about* than *The teacher wondered which book the students read quietly about*. This indicates that plausibility can not prevent a gap from being postulated. Similarly, in a cross-modal priming study, Hickok, Canseco-Gonzalez, Zurif, & Grimshaw, (1992) found evidence for reactivation of the filler immediately after the first verb even in sentences such as *Which bucket did the movie director from Hollywood persuade Bill to push?*, even though *bucket* is implausible as the direct object of *persuade*.

Although plausibility may not influence *whether* a direct object gap is postulated, it may nevertheless influence the reader's commitment to the resulting interpretation, and the probability that it will be cancelled even before further information is processed. Evidence for this comes from Boland, Tanenhaus, Garnsey, & Carlson (1995) who employed a "stop making sense task" in which subjects read sentences a word at a time and had to press a button as soon as they thought that the sentence had become implausible. They found that the rate of stop making

sense decisions at the verb in sentences such as *Which prize did the salesman visit ...* was higher than in sentences such as *Which movie did your brother remind ...* They argued that in the latter case, even though the filler was implausible as direct object, it could be interpreted as potentially filling a third argument role instead. This possibility reduced the reader's commitment to a direct object analysis, hence making a stop making sense judgement less likely.

Using a more naturalistic eye-movement tracking technique, Traxler & Pickering (1996) showed that the plausibility of a gap interpretation influences subsequent processing. They compared sentences such as (3) and (4).

(3) We like the city that the author wrote unceasingly and with great dedication about while waiting for a contract.

(4) We like the book that the author wrote unceasingly and with great dedication about while waiting for a contract.

They found that the filled-gap effect on *about while* was reduced when the filler was implausible as the direct object of the verb *wrote*, as in (3), compared to when it was plausible, as in (4), at least when total reading time was considered. At the same time, reading times showed the opposite pattern in the region *wrote unceasingly*, being slower in (3) than in (4). The interaction between plausibility and region shows that the plausibility of the filler as direct object of the verb was computed immediately at the verb, and when it was implausible this analysis was cancelled. This result not only demonstrates that native speakers employ a Filler Driven strategy for identifying gaps, but that they also immediately interpret the semantic plausibility of even a potential, but unconfirmed, filler-gap assignment. However, when the filler-gap relation is implausible it remains unclear whether this actually cancels the gap hypothesis, as would be predicted if there were a strong interaction between syntactic and semantic processing. Certainly the relatively fast reading of the disambiguating region at *about* might suggest that this is the case. But it is also possible that there is no revision until this is actually forced by the disambiguating region itself. This revision process may be less costly when the filler would have been implausible as the direct object of the verb than when it would have been plausible. From this perspective, the implausibility of the filler as direct object would not directly cancel the gap hypothesis, thereby preserving the modularity of syntax and semantics, but it would make it easier to abandon that analysis later on. Despite this ambiguity, manipulations of plausibility of this type do provide a valuable diagnostic of the immediacy of the semantic interpretation of putative filler-gap relationships during on-line processing, and it is for this reason that plausibility was manipulated in the present experiments.

Research issues in relation to gap processing in the second language

The above considerations raise a number of issues that are interesting to explore in the second language context:

(I) Are non-natives able to adopt a filler-driven strategy, as revealed by the filled-gap effect? This strategy appears to reflect a predictive, top-down, approach to parsing, as well as a high level of automaticity. Either or both of these may be difficult for non-natives to acquire, even at relatively high levels of proficiency. Instead, non-natives may adopt the apparently more logical, and more conservative, Gap as Last Resort strategy, only positing a gap when the current word would be otherwise uninterpretable.

(II) Is the ability to use a filler-driven strategy affected by the nature of the speaker's first language (L1)? In the present study we compared processing of English wh-constructions by advanced German, Chinese, and Korean learners of English. German forms wh-questions in a

similar fashion to English[1] by positing a trace that is co-indexed with a filler.[2] In Mandarin Chinese and Korean, the wh-phrase appears in the same position in questions as in declaratives, thereby eliminating the need for a wh-trace. For example, the question *Which car did the tourist buy?* would have the word order 'tourist buy which car?' in Chinese, and 'tourist which car bought ' in Korean. A comparison between these different L1 groups will indicate to what extent the existence of English-like processing routines in the L1 influences the extent to which a native-like parsing strategy can be acquired in the L2.

However, it should be noted that it is not the case that displacement of objects never occurs in either Chinese or Korean. Both languages allow for topic movement (Huang, 1984), so that in Chinese it is possible to say, for example 'Tea, I like', and in Korean, the possibility of scrambling allows for a variety of word orders, even an English-like word order in questions, as in 'Which car tourist bought'. Hence the difference between German and English on the one hand, and Chinese and Korean on the other, relates specifically to the existence of a trace in the surface structure of wh-questions.

(III) Will natives and non-natives differ in the their sensitivity to plausibility constraints? Whilst the filled gap effect provides a diagnostic of the use of a Filler Driven strategy, any modulation due to plausibility will indicate the extent to which hypothetical filler-gap relations are interpreted semantically during on-line processing. Just & Carpenter (1992) argue that in individuals with relatively low working memory capacity, initial syntactic parsing decisions are immune from semantic influences, whereas those with high working memory capacity show evidence of dynamic interactions between syntax and semantics. Similarly, Pearlmutter & MacDonald (1995) demonstrated that readers with poor performance on Just & Carpenter's reading span task (taken as a measure of working memory capacity) were relatively insensitive to the plausibility of alternative syntactic analyses. They also showed that this was a feature of on-line, rather than off-line processing. However they attributed these effects to differences in language experience, rather than working memory capacity. Non-native speakers are likely to have a lower working memory capacity in the second language than a native speaker, as well as more limited language experience.[3] One may therefore expect them to have more difficulty computing the plausibility of a potential gap structure during on-line processing.

However, in the second language context, the nature of the individual's first language may be a relevant factor. Given the lack of syntactic and morphological cues in Chinese, one might regard the Chinese native mode of processing to be heavily dependent upon the evaluation of plausibility in relation to the general discourse. In contrast, morphologically marked case

---

[1] This only applies to the extraction of arguments. Preposition-stranding is ungrammatical in German.

[2] Indeed, there is evidence for use of the filler-driven strategy by Dutch speakers in processing equivalent wh-gaps in Dutch (Frazier, 1987) despite obvious syntactic differences between Dutch and English (i.e., strictly verb-second in main clauses and verb-final in subordinate clauses in Dutch but not in English) which might at first sight suggest the non-applicability of the filler-driven strategy in Dutch. On the basis of the similarites between Dutch and German one might expect Germans to also operate a filler-driven strategy in German.

[3] In fact, no published studies have compared working memory for second language learners and native speakers of the same language. Osaka & Osaka (1992) and Harrington & Sawyer (1992) compared reading spans in the same individuals' L1 (Japanese) and L2 (English) and did not find a difference. In contrast, Osaka, Osaka, & Groner, (1993) found higher reading spans in L1 German than in L2 French. However, for the languages involved in these studies it is very difficult to equate the item characteristics. Walter (2000) compared reading spans in both French and English for French learners of English using carefully matched items in the two versions of the test. She found that reading spans were significantly higher in the L1 (French) than the L2 (English) tests at both intermediate and advanced levels of proficiency (the difference was larger at the intermediate than the advanced level). In Experiment 1 we measured the reading spans of all subjects in English, thereby enabling comparisons between natives and non-natives on the same test.

(either on nouns or wh-words) in German unambiguously marks the thematic role of the wh-phrase, making evaluation of plausibility unnecessary even for structures that would be temporarily ambiguous in English. Similarly, in Korean, although scrambling can result in flexible word order, subject and object markers generally make the grammatical roles of the sentence constituents clear without having to invoke plausibility constraints. From this perspective, one might predict that if general processing strategies transfer to L2 then Chinese would be more sensitive to plausibility constraints in on-line processing than Germans or Koreans. This might reflect transfer of the relative weight given to different cue domains, as predicted by the Competition Model (Bates & MacWhinney, 1989).

Previous L2 research

Few studies have examined on-line sentence processing in non-natives, but those that have tend to point more to similarities between natives and non-natives rather than to differences. There is good evidence that advanced non-natives are subject to garden-path phenomena in much the same way as natives (Frenck-Mestre & Pynte, 1997; Hoover & Dwivedi, 1998; Juffs, 1998; Juffs & Harrington, 1995; Juffs & Harrington, 1996). The studies of Hoover & Dwivedi (1998) and Juffs & Harrington (1995) are of particular relevance because they deal with structures containing empty categories.

Juffs & Harrington (1995) investigated on-line processing of wh-questions by advanced Chinese learners of English. They found that, compared to natives, Chinese learners have particular difficulty processing subject extractions of the type *What does the man think crashed into the car?*, compared to object extractions such as *What does the man think the car crashed into?*. Specifically, subject extractions led to slower reading times in the region following the second verb, *crashed*, than object extractions. Following Pritchett (1992) they assumed that in both types of sentence the wh phrase is initially assigned the theme role of *think*; that is, that the readers were utilizing the filler-driven strategy. The difference between the two types of sentence over the following region arose because of the greater cost of reanalysis in the subject extraction sentence. However, exactly the same outcome would be predicted if the readers were utilizing the gap as last resort strategy. In the case of the subject extraction, a gap is forced by the otherwise ungrammatical sequencing of the two verbs *think crashed*, and the slower reading time on the second verb could reflect the cost simply of deriving the gap analysis. In the object extraction condition no gap is postulated until the end of the sentence, and so reading times in the region following *think* would be relatively fast.

On the other hand, there is some evidence that Juffs & Harrington's (1995) Chinese subjects were showing a filled gap effect even for object extractions; the effect was just not as large as for subject extractions. Although these data were not subjected to an independent analysis, the Chinese appeared to show a slow-down in reading on the word after the first verb *think*, whereas the natives were faster at this position relative to the verb. However, it is not clear why this effect was absent in the natives, who on the basis of previous research ought also to have been employing the filler-driven strategy. One possibility is that since, in fact, one third of the items would have been ungrammatical with a gap after the first verb (e.g. *Who did Jane say her friend likes?*), then the natives were more sensitive to lexical and plausibility factors than the Chinese. This may be an indication that natives and non-natives differ in their ability to utilize plausibility constraints during parsing.

Hoover & Dwivedi (1998) examined processing of French clitics by advanced English learners of French. For a causative structure such as (5) it was found that both natives and learners read the verb *goûter* more slowly than in the non-clitic version in (6).

(5) Il le faisait tranquillement goûter avec son fromage préferé
He it had quietly tasted with his cheese favourite.
He had it be tasted quietly with his favourite cheese.
(6) Il faisait tranquillement goûter le vin avec son fromage préferé
He had quietly tasted the wine with his cheese favourite.
He had the wine be tasted quietly with his favourite cheese.

They argue that in the clitic structure the readers posit an empty category (in this case PRO) at *faisait* which is coindexed with the clitic *le*. In effect the clitic pronoun is interpreted as the direct object of the first verb, a hypothesis which has to be revised when the second verb, *goûter*, is encountered, since *le* is in fact the object of this verb. This kind of garden-path effect is therefore similar to the filled gap effect. Crucially, in this case it was obtained even in learners whose first language, English, does not contain clitics, or pre-verbal objects.

On the basis of previous research there is therefore good reason to expect that non-natives will exhibit a filled gap effect for wh questions. What is not so clear is whether effects of plausibility will be obtained. Neither of the above studies included a plausibility manipulation, so although there was evidence that non-natives carried out immediate and incremental syntactic processing, whether they performed immediate and incremental semantic and thematic processing was not addressed. Frenck-Mestre & Pynte (1997) found evidence for the on-line use of lexical information, but this took the form of transitivity preference, rather than plausibility as such. It remains unclear, therefore, whether non-natives fully interpret the meanings of the syntactic structures they compute on-line.

EXPERIMENT 1

Experiment 1 examined whether advanced nonnative learners of English exhibit a filled gap effect in on-line reading, and if so, whether they also show sensitivity to plausibility constraints. There were two critical conditions which are illustrated in Table 1.

Table 1. Example items from Experiment 1.

Filler plausible as object of verb (Plausible-at-V) Which girl did the man push the bike into late last night?
Filler implausible as object of verb (Implausible-at-V) Which river did the man push the bike into late last night?

The Filler-driven strategy predicts that on reading the verb *push* readers should initially interpret the filler, *girl/river*, as its direct object. When they encounter the following noun phrase, *the bike*, they should be forced to reanalyse the structure, and this should be evident as a slow-down in reading. The Gap as Last Resort strategy predicts that there should be no such slow-down because readers do not insert the filler into the surface structure until they are forced to do so by a sequence of words that would otherwise be ungrammatical. Since this only occurs after the preposition, at *late*, there should be no slow-down on the post-verbal noun phrase itself.

Strictly speaking, in order to be sure that any slow-down in the post-verbal region is a garden-path effect, reading times should be compared with a control condition in which the same verb-determiner-noun sequence occurs but with no gap (e.g. *They wondered whether the man would push the bike into the girl*). However, in the present experiment the additional manipulation of plausibility made this unnecessary, provided that some effect of plausibility was

obtained.[4] This is because any difference in reading times between the Plausible-at-V and Implausible-at-V conditions over the post-verbal noun phrase can only be because the filler was initially interpreted as the direct object of the verb, thereby simultaneously providing evidence of use of the Filler-driven strategy and sensitivity to plausibility.

One problem with investigating effects of plausibility during on-line processing is that it is difficult to distinguish an endemic inability to use such information from the failure to use it for other reasons, not least the fatigue caused by reading large numbers of unconnected sentences in a psycholinguistic experiment. When a null result (i.e. a failure to utilize plausibility) is potentially interesting it is important that this is obtained under conditions which truly motivate on-line reading for meaning. The present experiment therefore examined reading times when subjects were engaged in an on-line plausibility judgement task adapted from the stop making sense task used by Boland and colleagues (Boland et al., 1995). In this task subjects read sentences one word at a time in order to provide a measure of on-line reading times. They also had to press a button as soon as they thought that the sentence had become implausible. This methodology therefore provides two dependent measures: first, the rate of stop making sense decisions at various sentence positions, and second, assuming that these are generally low (because the critical sentences are all ultimately plausible) the reading time for sentences where no stop making sense decision was made prior to, or during, the critical region for analysis.

We recognise, however, that the task used here is unnatural. Subjects are forced to read one word at a time and, unlike in normal reading situations, they know that some of the sentences will be implausible. In the case of native speakers, filled-gap effects, and plausibility effects, have been obtained in eye movement tracking studies where the task was to answer yes/no comprehension questions (Pickering & Traxler, 1998; Traxler & Pickering, 1996). This suggests that the processes involved are not contingent on specific presentation or task conditions, and that they are automatically carried out in the course of sentence processing (for evidence of the comparability of reading patterns under word-by-word reading and eye movement tracking see Ferreira & Henderson, 1990). In the case of non-natives it is important to separate the issue of automaticity from that of whether the requisite knowledge and processes have been acquired. The purpose of Experiment 1 was to evaluate whether non-native speakers of English can, in principle, utilise a filler-driven strategy and exploit plausibility constraints during on-line processing in a task that would maximise the probability of such effects occurring. Whether their access to the relevant knowledge, and the ability to utilise it, is sufficiently automatic to produce the same effects in other reading situations is a separate issue.

All subjects in Experiment 1 performed a reading span test of the type developed by Daneman & Carpenter (1980). The aim of this was to see if any group differences in the use of plausibility constraints could be related to differences in working memory.

Subjects

There were 75 participants. A total of 18 subjects were tested in the native, German, and Chinese groups. These subjects were drawn from the student and post-doctoral populations of the University of Cambridge. All non-native students at Cambridge must have achieved the Cambridge Certificate of Proficiency in English or an equivalent EFL qualification, which ensures that they are linguistically functional in the academic environment. In addition 21

4 Pickering & Traxler (1998) point to problems in using control conditions which are syntactically different from the condition of interest. They suggest that manipulations of plausibility, such as that employed here, are a good way to investigate misanalysis processes because the relevant comparisons are made across virtually identical sentences.

Koreans were tested. They were undergraduates, graduate students and staff from the National University of Chonnam, South Korea, taking English courses in the Language Education Centre at advanced level. Each subject completed a questionnaire in order to obtain the following background information: Age, total amount of formal instruction in English in months (Instruction), time spent in an English-speaking country in months (Immersion), a subjective estimate of the average amount of time spent reading in English per day in hours (Reading). The biographical data for each group are summarized in Table 2 along with the results of the reading span test.[5] Scheffe tests showed that the reading span scores for the Koreans and Chinese were significantly smaller than those for the Germans and natives.

Table 2. Biographical data for the subjects who participated in Experiment 1

	N	Age	Instruction	Immersion	Reading	Reading Span
Koreans	21	26.7 (4.5)	77.14 (13.8)	3.9 (7.3)	9.9 (11.6)	44.0 (9.6)
Chinese	18	30.1 (7.4)	103.3 (35.0)	54.6 (66.2)	4.1 (2.2)	43.6 (17.2)
Germans	18	26.8 (3.9)	99.3 (23.6)	43.9 (41.1)	3.3 (1.6)	63.7 (7.9)
Natives	18	28.2 (5.2)				59.6 (9.8)

Standard deviations in parentheses.

Method

Materials

The critical materials were constructed around 16 core sentences containing an adjunct extraction. Two versions of each sentence were written, differing only with respect to the filler noun, as illustrated in Table 1. In the Plausible-at-V version the filler noun was plausible as direct object of the verb. In the Implausible-at-V version the filler noun was implausible as direct object of the verb. In both versions the correct interpretation of the sentence, in which the filler noun fills a gap after the preposition in the adjunct phrase, was plausible. The items are listed in the Appendix.

For the purposes of testing the filler-driven strategy it is merely necessary that the filler noun is plausible as direct object of the verb, and that it actually turns out to fulfil an alternative role later in the sentence. However, verbs that obligatorily take two internal arguments, such as *put*, were avoided because it is possible that the availability of the alternative role in the representation of the verb could reduce the garden-path effect. For fifteen out of the sixteen items the filler was clearly extracted from an adjunct phrase, but for one of the items (with the verb *buy*) it is possible that the *for* phrase corresponds to an optional internal argument (for example, this construction is dativizable).

---

[5] In this test, subjects read aloud sets of sentences presented one at a time on the computer screen. At the end of each set they were required to recall the final word of each sentence, and then answer a comprehension question about one of the sentences. Set size increased from two to six, and testing terminated when the subject could not recall more than 2 out of 5 sentence-final words at a particular set size. The scoring system combined recall and comprehension scores, weighted by set size. The data in Table 2 are the mean scores over set sizes one and two only. The maximum score is 75. The data for one Korean subject were unavailable. The Germans and natives did not differ when performance over all set sizes was taken into consideration, and scoring was in terms of the maximum set size at which 4 out of 5 of the sentence-final words could be recalled (with half a point for 3 out of 5). Using this scoring method the mean for the Germans was 2.92 (SD = 0.575) and that for the natives was 3.02 (SD = 0.795).

The 16 items were divided into two groups of 8 items each. Two presentation lists were constructed. In the first, the items from one of the item groups appeared in the Plausible-at-V condition and the items from the other group appeared in the Implausible-at-V condition. These assignments were reversed to form the second presentation list. Half of the subjects in each language group received each list.

A large number of filler items were also written, some of which constituted a sub-experiment which will not be reported here. There were 38 implausible sentences comprising the following: twelve implausible declaratives in which the direct object was implausible (e.g. *The rich family burned the key in their heater late last night*), two declaratives in which the adjunct was implausible, eight implausible argument extractions with a relatively short distance between filler and gap site (e.g. *Which shop did the criminal kill in the city yesterday evening*?), eight implausible argument extractions with a relatively long distance between filler and gap site (e.g. *Which bird did Janet's older brother Frank build at the end of the road last year*?), and eight implausible adjunct extractions (e.g. *Which desk did the secretary find the letter with earlier this morning*?). There were also 32 sentences in which there was no implausibility: twenty declaratives, six short argument extractions, and six long argument extractions. The sentences for each presentation list were presented in three different pseudo-random orders, with the proviso that no two sentences of the same type should occur consecutively.

There were 19 practice sentences of which 10 contained an implausibility. Ten sentences were wh questions.

Procedure

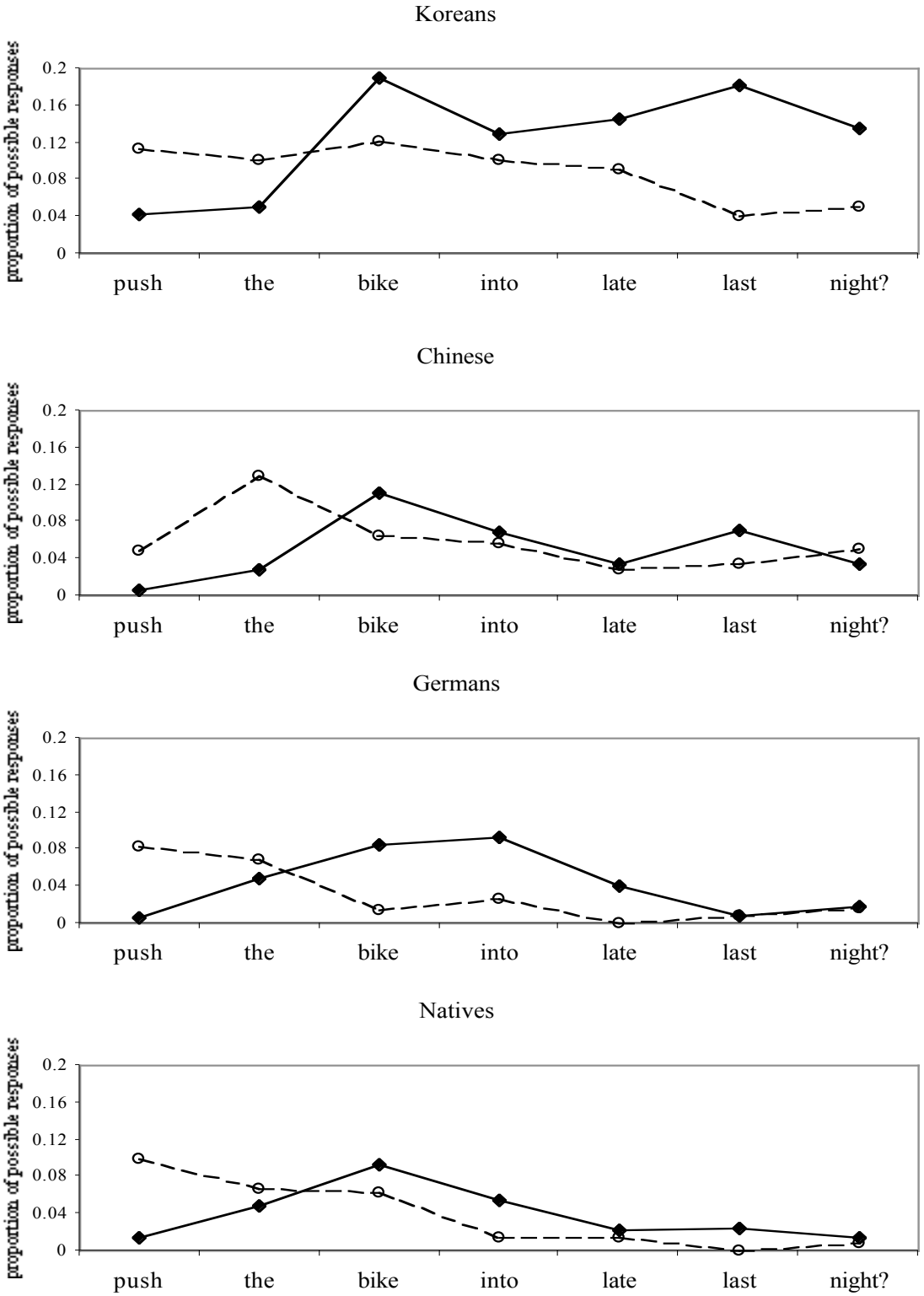
Sentences were displayed one word at a time on a computer screen, all words appeared in the same position half way down the screen near to the left-hand edge, and subjects changed each word into the next by single-clicking the left mouse button with their dominant hand. They were also instructed to respond as soon as they thought that the sentence had stopped making sense by pressing the space bar with their non-dominant hand. This response also had the effect of changing the current word into the next word. After any stop making sense decision the subject continued reading the remainder of the sentence by pressing the mouse button. All times between presentation of a word on the screen and a response were measured to millisecond accuracy by the software.

Results

Stop making sense decisions

The 'stop making sense' decisions were analysed in terms of the position in the sentence at which they were made. Of most interest were the responses made at and following the verb, that is, at positions defined as: verb, determiner, noun, preposition, preposition +1, preposition + 2, and remainder of sentence (e.g., *push the bike into late last night*). In order to compare response rates at individual positions, the number of responses made at earlier positions had to be taken into account. Response rates at each position were calculated relative to the number of possible responses that could have occurred at that position (only one stop making sense decision was coded for each sentence, and in the very rare cases where more than one response was made, only the earliest was counted). For example, a subject who responded at the verb in two of the eight sentences in the Plausible-at-V condition and at the noun in a further two sentences would have a response rate of 0.25 (2/8) at the verb and 0.33 (2/6) at the noun. The resulting response rates for each condition and each group are displayed in Figure 1. A similar analysis was performed on the data when organized by items.

Figure 1. Stop making sense decision rates in Experiment 1.
Solid lines: Plausible-at-V condition (*Which girl did the man ..*)
Dashed lines: Implausible-at-V condition (*Which river did the man ..*)



An analysis of variance was performed on the data organized by subjects in which Language and Presentation List were between-subjects' factors, and Plausibility (Plausible-at-V or Implausible-at-V) and Position (with 7 levels) were within-subjects' factors. An analysis was also performed on the data organized by items in which Language and Position were within-items' factors, and Item Group and Plausibility were between-items' factors[6].
There were main effects of Language, $F1(3,67) = 8.35$, $p<0.001$, $F2(3,84) = 13.48$, $p<0.001$. Although there was no main effect of Plausibility, F1 and F2 both < 1.0, there was an interaction between Plausibility and Position, $F1(6,402) = 6.98$, $p<0.001$, $F2(6,168) = 6.76$, $p<0.001$. There was no interaction between Language, Plausibility, and Position, F1 and F2 both <1.0. The overall interaction between Plausibility and Position arises because for all groups there is a clear cross-over between the conditions at the determiner and noun, e.g. *the bike*. For example, the Korean subjects were more likely to make stop making sense decisions after reading *Which river did the man push* .. than after *Which girl did the man push* .. . This merely reflects the difference in plausibility of the filler as direct object of the verb. However, if subjects read through this region without making a decision, they were then more likely to make a stop making sense decision after *Which girl did the man push the bike* ... than after *Which river did the man push the bike* ... . It is the difference in response rates at and following the noun which are of most interest, because even though neither of these fragments are complete in themselves, subjects were deciding that *Which girl did the man push the bike* ... had become irrevocably implausible more often than *Which river did the man push the bike* ... . The data for the positions at and following the noun were submitted to an analysis of variance in which Position had 5 levels. The effect of Plausibility was significant, $F1(1,67) = 13.03$, $p<0.001$, $F2(1,28) = 9.35$, $p<0.01$, and there was no interaction between Language and Plausibility, $F1(3,67) = 1.56$, $F2 < 1.0$.

In order to assess whether the significance of the above effects was due to the inclusion of the natives, the analysis of response rates following the noun was repeated with the native group excluded. Once again there was a main effect of Plausibility, $F1(1,51) = 10.06$, $p<0.01$, $F2(1,28) = 10.58$, $p<0.001$, and no interaction between Language and Plausibility, $F1(2,51) = 1.55$, $F2 < 1.0$. Further analyses of variance on each language group showed that the effect of plausibility in this region was significant for the Koreans, $F1(1,19) = 4.73$, $p<0.05$, $F2(1,28) = 6.88$, $p<0.05$, for the Germans, $F1(1,16) = 14.21$, $p<0.01$, $F2(1,28) = 4.63$, $p<0.05$, for the natives by subjects but not by items, $F1(1,16) = 6.86$, $p<0.05$, $F2(1,28) = 2.54$, and for the Chinese on neither analysis, $F1(1,16) = 2.07$, $F2(1,28) = 2.02$.

The effect of plausibility in the verb-determiner noun region is also of relevance since this indicates the extent to which subjects were computing the plausibility of the filler-gap relationship during on-line processing. Analyses of variance showed that there was a main effect of Plausibility in this region, $F1(1,67) = 29.96$, $p<0.001$, $F2(1,28) = 6.43$, $p<0.05$. Independent analyses of each language group showed that the effect of plausibility was significant on the subjects' analysis for all groups, only significant on the items' analysis for the Chinese, and approaching significance on the items' analysis for the remaining groups: natives, $F1(1,16) = 6.49$, $p<0.05$, $F2(1,28) = 3.94$, $p = 0.057$; Germans, $F1(1,16) = 6.95$, $p<0.05$, $F2(1,28) = 3.07$, $p = 0.091$; Koreans, $F1(1,19) = 6.61$, $p<0.05$, $F2(1,28) = 3.78$, $p = 0.068$; Chinese, $F1(1,16) = 10.96$, $p<0.01$, $F2(1,28) = 7.50$, $p<0.05$.

In general, then, these results show clear effects of plausibility on stop making sense decisions. Higher response rates at, and immediately following, the verb in the Implausible-at-V condition were to be expected, although the failure to achieve significance on the items' analysis

---

[6] Plausibility was a between-items' factor because the sentences in the Plausible- and Implausible-at-V conditions contained different fillers and in some cases the prepositions were also different

shows that this effect is not equally strong for all items. What is perhaps surprising is the higher rate of responding in the Plausible-at-V condition at and following the noun. This provides initial evidence for an influence of plausibility on the filled gap effect. Although the pattern of data for the Chinese was similar to the other groups, the effect of plausibility at and following the noun failed to reach significance for this group.

Reading times

For the purposes of the reading time analysis the verb, determiner, and noun were defined as constituting the critical region (e.g. *push the bike*). Outlying reading times (both high and low) were identified by subject and condition using the sample kurtosis as discordancy test (Barnett & Lewis, 1994, p. 231). These were replaced by the next highest reading time for that subject in that condition and at that position. All data from trials on which a stop making sense decision was made prior to the end of the critical region were then excluded. Hence, the reading time data that are reported here are just for those trials on which a stop making sense decision was not made before or during the critical verb-determiner-noun region. For some subjects this meant that there were too few trials in one or both of the conditions to make a reaction time analysis tenable. It was decided to remove subjects who made more than 4 (out of a possible 8) stop making sense decisions in either condition. This resulted in the loss of data for 5 Koreans and one native.

Mean response times for each condition were submitted to a preliminary analysis in order to identify any subjects with atypical reading patterns[7]. The analysis was performed on the percentage change in reading time from the subject noun to the following verb, from the verb to the determiner, and from the determiner to the noun. Separate hierarchical cluster analyses were then performed on the mean percentage change scores for each group in each condition (these analyses employed squared Euclidean distance and the weighted pair-group average linkage rule). We looked for cases where the final (and largest) step in the amalgamation schedule resulted in a cluster of one or two cases. These were then excluded from the analysis of both conditions. This procedure resulted in the exclusion of one Korean, three Chinese, two Germans, and one native. The mean percentage of the total linkage distance accounted for by these atypical cases was 48% (*SD* = 10%). The reading patterns for these excluded subjects will be discussed further below. First we will report the results of the analyses based on the data from the remaining 15 Koreans, 15 Chinese, 16 Germans, and 16 natives. The reading times for each language group in the two conditions are shown in Figure 2.

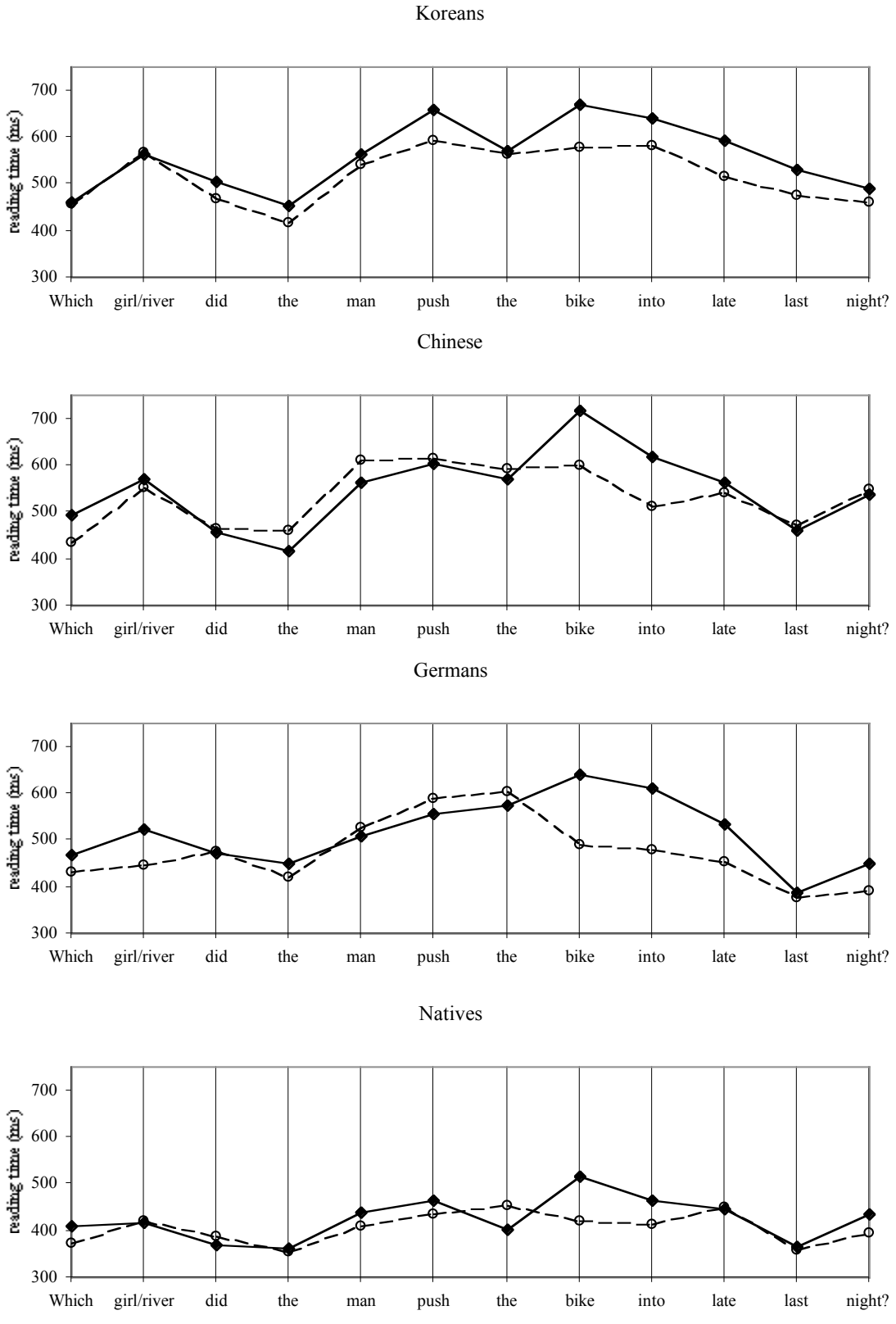
Analyses of variance by both subjects and items were conducted on the reading times in the critical region. For the analysis by subjects Language and Presentation List were between-subjects factors, and Plausibility and Position (verb, determiner, noun) were within-subject factors. For the analysis by items Item Group was a between-items factor, and Language, Plausibility, and Position were within-items factors[8]. There was a main effect of Language, $F1(3,54) = 3.19$, $p<0.05$, $F2(3,42) = 39.45$, $p<0.001$. The main effect of Plausibility was only significant on the subjects' analysis, $F1(1,54) = 7.05$, $p<0.01$, $F2(1,14) = 3.42$. There was a significant interaction between Plausibility and Position, $F1(2,108) = 12.22$, $p<0.001$, $F2(2,28) = 12.49$, $p<0.001$. The interaction between Language, Plausibility, and Position was not significant, F1 and F2 both < 1.0.

---

[7] This was felt to be necessary because a minority of subjects showed very large increases in reading time at specific positions. Since the aim of the analysis was to ascertain the reading pattern for the majority of subjects it was felt that the influence of subjects with atypical reading patterns should be curtailed.

[8] Plausibility was a within-items factor because the words in the verb-determiner-noun region were identical in the two conditions.

Figure 2. Mean reading times (msecs) in Experiment 1.
Solid lines: Plausible-at-V condition (N1 = *girl*)
Dashed lines: Implausible-at-V condition (N1 = *river*)



In order to gauge whether the significance of the above effects was due to the inclusion of the natives, a similar analysis was performed when they were excluded. The main effect of Language now disappeared, $F1 < 1.0$, $F2(2,28) = 2.13$. The main effect of Plausibility was again only significant on the subjects' analysis, $F1(1,40) = 5.29$, $p<0.05$, $F2(1,14) = 4.13$. There was a significant interaction between Plausibility and Position, $F1(2,80) = 7.85$, $p<0.001$, $F2(2,28) = 8.90$. $p<0.001$, and no interaction between Language, Plausibility, and Position, F1 and F2 both < 1.0.

In order to determine the consistency of the Plausibility by Position interaction, separate analyses of variance were conducted on the reading times in the critical region for each language. The interaction between Plausibility and Position was significant for the natives, $F1(2,28) = 7.15$, $p<0.01$, $F2(2,28) = 8.09$, $p<0.01$, and Germans, $F1(2,28) = 4.39$, $p<0.05$, $F2(2,28) = 8.41$, $p<0.01$. For the Chinese this interaction was only significant by items, $F1(2,26) = 2.53$, $F2(2,28) = 5.26$, $p<0.05$, and for the Koreans it was not significant on either analysis, $F1(2,26) = 1.83$, $F2(2,28) = 1.11$. However, given that in all groups the difference between conditions was localized at the noun, the effect of plausibility was tested solely at the noun position in the Chinese and Koreans. For the Chinese this difference was significant on both analyses, $F1(1,13) = 8.85$, $p<0.05$, $F2(1,14) = 8.84$, $p<0.05$, and for the Koreans it was significant on the subjects' analysis, $F1(1,13) = 9.14$, $p<0.01$, and approached significance on the items' analysis, $F2(1,14) = 4.09$, $p= 0.062$.

Figure 2 shows that only the natives and Koreans appeared to show any effect of plausibility prior to the post-verbal noun. The Koreans showed longer reading times on the verb in the Plausible-at-V condition, but this effect was not significant, $F1(1,13) = 1.21$, $F2(1,14) = 2.46$. In any case there seems no reason to expect a slow down at this position in this condition. The natives showed longer reading times at the determiner in the Implausible-at-V condition than the Plausible-at-V condition, whereas at the following noun this difference was dramatically reversed. The difference in reading time at the determiner was found to be significant, $F1(1,14) = 6.41$, $p<0.05$, $F2(1,14) = 19.18$, $p<0.001$, as was that at the noun, at least on the analysis by subjects, $F1(1,14) = 6.41$, $p<0.05$, $F2(1,14) = 4.27$, $p = 0.058$. Therefore, it appears that in the natives the interaction between Plausibility and Position was driven by a more dramatic shift in reading pattern over the critical region than in the non-native groups.

Finally, the data for the subjects who were excluded on the basis of their divergent reading profiles will be considered. These subjects were separated out by the cluster analysis because they showed very large peaks in reading time at points in the sentences which were in nearly all cases atypical compared to the rest of their group. The positions of these peaks are shown in Table 3, along with the subject codes. Since the cluster analysis took into account the pre-verbal noun, this position is also included. The increase in reading time with respect to the preceding word is shown.

Table 3. Positions of peaks in reading time for the subjects identified as displaying divergent reading profiles.

	Plausible-at-V				Implausible-at-V			
position	N2 (girl)	V (push)	det (the)	N3 (bike)	N2 (river)	V (push)	det (the)	N3 (bike)
subject code	C4 (+1005)	N2 (+609)	G15 (+522)	C18 (+908)	N2 (+956)	G8 (+862) K21 (+628)		C4 (+641) C5 (+934)

Note: K = Korean, C = Chinese, G = German, N = Native

Considering first the Plausible-at-V condition, C18 peaks at N3 and therefore conforms to the general tendency to slow down at this position. The peak for G15 at the determiner could be because this is taken early evidence that the filler is not the direct object of the verb. However, as the profile for the rest of the group shows, such early sensitivity is atypical of all groups, including the natives. Why C4 should have peaked at N2 is not clear, unless there was some problem with the lexis at this position. In the Implausible-at-V condition, N2 showed a similar tendency, but again there is no obvious explanation for this. The peaks for G8 and K21 at V are to be expected given that the filler was implausible as direct object, and the increased rate of stop making sense decisions at this position (see Figure 1). It is perhaps surprising that more subjects did not peak at this position (see Figure 2). The peaks for C4 and C5 at N3 were atypical, both for their group and the sample as a whole. These subjects showed an extreme tendency to slow down on the post-verbal noun even when the filler was implausible as the direct object of the verb. For C4 the effect of plausibility was evident on the following preposition, however, where reading times were much slower in the Plausible-at-V condition than the Implausible-at-V condition. This subject therefore appeared to show a delayed sensitivity to plausibility. C5 showed a very similar reading pattern to C4 in the Implausible-at-V condition, but did not show plausibility effects in the following region.[9]

Discussion

The stop making sense decision and reading time data provide evidence that for all groups of subjects the plausibility of the filler as direct object had an impact on task performance. All but the Chinese subjects showed a greater tendency to make stop making sense decisions at and following the post-verbal noun in the Plausible-at-V condition, and all groups showed longer reading times on the post-verbal noun in this condition (barring the two

[9] We looked for explanations of the divergent behaviour of C4 and C5 in terms of other measures that were available for these subjects. C4 and C5 had relatively low scores on the working memory test, and their scores were outside the range of the natives and Germans. In order examine the possible effect of low working memory, those Korean and Chinese subjects with working memory scores outside the range of natives and Germans were examined separately. If C4 and C5 were included in this sample (n = 12) then there was no effect of Plausibility at the post-verbal noun, $F < 1.0$, whereas there was a very strong effect for the 12 Chinese and Korean subjects with the highest working memory scores, $F(1,11) = 16.33$, $p<0.01$. If C4 and C5 were excluded, then the effect of plausibility for the low working memory group increased, and was of marginal significance, $F(1,9) = 4.92$, $p = 0.054$. Therefore, there is some evidence that low working memory might reduce sensitivity to plausibility, but the contrast with the high working memory group is heavily influenced by just two subjects.

exceptional Chinese discussed above, only one of whom failed to show a plausibility effect completely). In addition, all subject groups made more stop making sense decisions at the verb in the Implausible-at-V condition than the Plausible-at-V condition, although there was little evidence for an effect of plausibility on verb reading times for trials where a stop making sense decision was not made.

These results show simultaneously that all groups of subjects were (a) analysing the filler as direct object of the verb, and (b) computing the plausibility of the filler as direct object. Thus, they were all performing as predicted by the Filler Driven strategy, and, furthermore, were actively interpreting the meaning of the potential filler-gap relationship during on-line processing. It is remarkable that evidence for these processes was obtained even for Chinese and Koreans, whose L1 differs radically from English in relation to question formation, and despite the fact that their working memory scores were significantly lower than those for the other groups. Thus, there was no evidence for a relationship between sensitivity to plausibility constraints and working memory capacity.

For all three non-native groups it is clear that the post-verbal noun phrase was easier to process when the filler was implausible as direct object of the verb. However, there was in fact very little evidence that this was because reanalysis had taken place prior to the occurrence of the noun itself.[10] If plausibility information had been used to immediately rule out the gap hypothesis, then one might have expected to see longer reading times on, or immediately following, the verb in the Implausible-at-V condition than the Plausible-at-V condition. However, no significant effect of plausibility was apparent until the noun in the non-native groups. The implausibility of the filler as direct object of the verb did not appear to immediately trigger syntactic reanalysis.

What kind of reanalysis process would only be evident at the post-verbal noun? The garden-path sentences used in the present experiment do not demand such extensive revision of phrase structure as other types of garden-path sentence, such as reduced relatives for example. Consider (7) and (8).

(7) [Which girl]$_i$ did the man push [t_i]?

(8) [Which girl]$_i$ did the man push the bike into [t_i]?

If it is assumed that these structures are built incrementally during comprehension, then the only additional processing required to construct (8) is to detach the post-verbal trace in (7), and its associated filler *girl*, from the object role, replace it with the lexical NP *the bike,* and then add the prepositional phrase as an adjunct to the VP. No actual revision of the phrase structure is required to accommodate the adjunct.[11] The fact that in non-natives the influence of plausibility is first apparent at the noun suggests that the process of detaching the trace/filler does not occur until this position, even though implausibility (in the Implausible-at-V condition) and the post-verbal determiner might be thought to provide evidence that reanalysis is necessary. In fact, the reading time at this point could in fact be regarded simply as a reflection of the cost of substituting one argument for another in the developing thematic structure of the sentence (i.e. substituting *bike* for *girl* as object of *push*). When the filler is plausible as direct object of the verb then it resists displacement by the post-verbal noun as theme of the verb, and reading times

---

10 Of course it is always possible that the peak in reading time at the noun in the Plausible-at-V condition is a spill-over effect from processing difficulty encountered at the determiner. However, in the absence of any difference between conditions at the determiner itself there seems no reason to assume that this is the case.

11 This is not to say that the requisite syntactic reanalysis is not problematic. According to Pritchett (1992) the fact that the trace has to be moved out of the theta-domain of the verb ought to cause processing difficulty. However, in the absence of a no-gap control condition it is not possible to gauge the difficulty of syntactic reanalysis in this experiment, only the impact of plausibility on the reanalysis process.

are relatively slow, or the subject may even judge the sentence to have stopped making sense. When it is implausible then it is more readily displaced, leading to faster reading times, and a lower likelihood of a stop making sense judgement. In other words, the filled gap effects evident in the non-natives may reflect the cost of thematic, rather than syntactic, reanalysis.

Only the natives showed any significant effect of plausibility prior to the point of argument substitution. Reading times at the determiner were significantly slower in the Implausible-at-V than the Plausible-at-V condition. It is reasonable to suppose that this reflects heightened sensitivity to the syntactic cue provided by the determiner (which signals that an unexpected noun phrase is coming) in this condition. Both semantic and syntactic information could be combining to detach the filler from the direct object and theme role prior to it being forcibly removed by the following noun. Thus, there is evidence that the natives can use a combination of syntactic and semantic information to undertake reanalysis more readily than the non-natives, prior to the forcible eviction of the filler by the post-verbal noun.

The present experiment therefore suggests that natives and non-natives do not differ in the kind of parsing strategy they use. The results support the notion of the Filler-driven strategy as a characteristic of first pass syntactic processing which applies regardless of plausibility constraints. There was no evidence that implausibility triggered immediate reanalysis, since there was no effect of plausibility at the verb in any group, and not even at the following determiner in the non-native groups. These results would appear to support the broad class of parsing models which assume that initial parsing decisions are made according to purely syntactic criteria (Gorrell, 1995; Pritchett, 1992), whilst extending this notion to non-native parsing. However, it has been argued that plausibility does influence the reanalysis process. Whilst both natives and non-natives showed clear evidence of sensitivity to plausibility in this sense, there was evidence that the natives initiated reanalysis more readily than the non-natives. Experiment 2 explores this possibility in the context of an off-line version of Experiment 1.

EXPERIMENT 2

One rather surprising aspect of the results of Experiment 1 was the effect of plausibility on stop making sense judgements at and following the post-verbal NP. It appears that occasionally the filler resisted displacement as theme of the verb to the extent that the sentence was judged to have stopped making sense. But is this 'mistake' merely a result of forcing subjects to make plausibility judgements during rapid self-paced reading? Given sufficient processing time, would the post-verbal noun be successful in displacing the filler as theme, regardless of its plausibility? In fact, this is one respect in which differences between language groups may emerge because reanalysis difficulties might be more persistent in some groups than in others.

In this experiment a pencil-and-paper version of the stop making sense task was administered to new groups of Korean, Chinese, German, and native subjects. The test simply consisted of the running lists from Experiment 1, and the subjects were required to indicate which sentences did not make sense, and if so, which word caused the problem.

Subjects

A total of 72 subjects participated. There were eighteen Koreans, Chinese, Germans, and natives respectively. The Koreans were drawn from the same population as in Experiment 1. The Chinese were first year undergraduates studying English at Beijing Foreign Studies University, and were following a course in intensive reading. The Germans were university students in

Germany at the time of testing. They were all students of Anglistik (English language, literature, and pedagogy) at the University of Augsburg. The natives were all graduates employed in local companies at the time of testing. The biographical data for these groups appears in Table 4. The characteristics of the Korean sample are similar to those in Experiment 1. The Germans are similar in all respects except the present Germans had far less immersion. The Chinese clearly have less experience with English than the subjects in Experiment 1. They are younger, have received less instruction and immersion, but claim to read for more hours per week.

Table 4. Biographical data for the subjects who participated in Experiment 2.

	N	Age	Instruction	Immersion	Reading
Koreans	18	27.1 (5.6)	78.0 (13.8)	7.9 (15.2)	9.1 (6.5)
Chinese	18	19.0 (0.5)	73.6 (18.0)	0.1 (0.4)	7.9 (6.8)
German	18	26.2 (4.0)	97.0 (13.6)	9.08 (6.8)	5.9 (3.5)
Native	18	31.8 (6.1)			

Method

The same six presentation lists were used as in Experiment 2 (two rotations of items, with 3 random orders each). The sentences were simply typed one sentence to a line on paper, and in the same order that they had appeared in the original presentation lists. The instructions stated that the subjects were to mark the sentences as either being 'okay' or as ungrammatical/implausible, and in the latter case they should put a circle around the word that made them think that the sentence was not okay. Eight examples were provided with the correct answers. One of these was ungrammatical because it contained an unnecessary preposition (*The beautiful woman drank the wine on last night*) and six contained semantic implausibilities (e.g. *The tall man sat on the universe*). Grammaticality was mentioned in the instructions for this task because we wanted the subjects to respond simply on the basis of whether they felt that the sentence was 'wrong', and not to worry about the distinction between ungrammaticality and implausibility. The instructions ended with the following note: "Read each sentence *at a normal speed once*, and make your decision immediately. *Do not read a sentence over and over again* before you make your decision about it. We are interested in your spontaneous responses."

Results

Four position categories were used for scoring: first noun (N1), verb (V), post-verbal noun (N3), and preposition (Prep). For example, for the item *Which girl did the man push the bike into late last night?* the scored positions were *girl*, *push*, *bike*, and *into*. The mean percentages of words circled at each position are shown in Table 5. There were no responses at the second noun (e.g. *man*), and although a small number of responses were made after the preposition, these showed no effect of plausibility, and seemed to reflect problems with temporal expressions.

Table 5. Mean percentages of words circled at each position in Experiment 2.

		N1	V	N3	Prep	N1+N3	Total
Koreans	Plaus-at-V	9.0	1.4	18.7	13.9	27.7	43.1
	Implaus-at-V	4.9	2.8	4.9	7.6	9.8	20.1
	difference	4.1	-1.4	13.8***	6.3	17.9***	23***
Chinese	Plaus-at-V	7.6	2.1	2.1	6.9	9.7	18.7
	Implaus-at-V	0	0	0	8.3	0	8.3
	difference	7.6**	2.1	2.1	-1.4	9.7**	10.4**
Germans	Plaus-at-V	3.5	4.2	6.9	7.6	10.4	22.2
	Implaus-at-V	0	4.2	0.7	0.7	0.7	5.6
	difference	3.5 (*)	0	6.2+ (**)	6.9+	9.7*	16.6**
Natives	Plaus-at-V	0.7	0	1.4	0.7	2.1	2.8
	Implaus-at-V	1.4	0.7	1.4	1.4	2.8	4.9
	difference	-0.7	-0.7	-0.1	-0.7	-0.7	-2.1

Note: difference = Plausible-at-V minus Implausible-at-V
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1. P values that are significant only on the items analysis are shown in parentheses, otherwise p values refer to both subjects and items analyses.

From the perspective of garden-path effects, and potential problems with argument substitution at the verb, the critical positions for analysis are the filler noun (N1) and the post-verbal noun (N3) since these are the nouns that are effectively competing for the object role. Responses at the preposition could simply reflect judgements about the acceptability of the preposition itself (e.g. the use of *for* in *Which meal did the chef cook the meat for this afternoon?*). On the basis of Experiment 1, judgements of implausibility might be expected at the verb in the Implausible-at-V condition, but there was clearly no evidence for such an effect in this task. The statistical analyses therefore focussed on the filler and post-verbal nouns (N1 and N3 respectively), and the total percentage of responses at these positions is given as N1+N3 in Table 5.

Analyses of variance were performed by both subjects and items. For the subjects analysis Language was a between-subjects factor, and Plausibility and Position (N1 and N3) were within-subjects factors. For the items analysis Item Group and Plausibility were between-items factors, and Language and Position were within-items factors.[12] There were main effects of Language , $F1(3,64) = 8.51$, $p<0.001$, $F2(3,84) = 28.11$, $p<0.001$, and Plausibility, $F1(1,64) = 37.21$, $p<0.001$, $F2(1,28) = 17.17$, $p<0.001$. There was also an interaction between Language and Plausibility $F1(3,64) = 6.48$, $p<0.001$, $F2(3,84) = 7.75$, $p<0.001$. As can be seen from the total of N1 and N3 responses in Table 5, the effect of plausibility was greatest for the Koreans, intermediate for the Chinese and Germans, and non-existent for the natives. There was also a three-way interaction between Language, Plausibility, and Position, $F1(3,64) = 2.76$, $p<0.05$, $F2(3,84) = 3.09$, $p<0.05$. This arises because the effects of plausibility (indicated by the difference scores in Table 5) are differently distributed over N1 and N3 in the three non-native groups. In order to explore this interaction further, the effect of plausibility at each position (this time including the preposition) and for each language group was evaluated using individual analyses of variance. Analyses were also performed on the data collapsed over position. Significant plausibility effects are indicated in Table 5. For the Koreans, the effect of plausibility is concentrated at the post-verbal noun phrase, N3 (that is, the point in the sentence at which

[12] Plausibility was a between-items factor because different filler nouns occurred in the two conditions

garden-path effects were evident in Experiment 1). For the Chinese the effect is concentrated at the filler noun, N1, whilst for the Germans it is weakly present at both N1 and N3, although the effects are not consistent over subjects. The mean plausibility effects were significant for all non-native groups, both when calculated over all positions, and over N1 and N3 alone.

Discussion

Although the non-natives' overall response rate was generally quite low, they made more responses to items in the Plausible-at-V condition than the Implausible-at-V condition. In contrast, the native speakers made very few responses to items in either condition. Therefore, it appears that the Plausible-at-V items had a tendency to seem unacceptable to the non-natives.

Because this was an off-line task, the precise word that subjects chose to circle does not offer such direct evidence of underlying processing difficulties as reading times. For example, for the Plausible-at-V item *Which meal did the woman cook the meat for during the afternoon?* non-native subjects circled *meal* (3 subjects), *cook* (1 subject), *meat* (4 subjects), and *for* (3 subjects). In contrast, for the matched Implausible-at-V item *Which pan did the woman cook the meat in during the afternoon?* subjects circled *meat* (2 subjects) and *for* (2 subjects), but no subject circled *pan*. In the discussion to Experiment 1 it was suggested that when reading Plausible-at-V items, subjects may have difficulty substituting the post-verbal noun (*meat*) for the filler noun (*meal*) as direct object of the verb (*cook*). The response to such a problem might be to circle the post-verbal noun (*meat*) as being unsatisfactory, since this is perceived as not fitting into the sentence, or as being in some sense superfluous. The Koreans showed a strong tendency to react in this way, and a weaker trend was also evident in the German data (although for the latter the effect, while consistent over items, was confined to a few subjects). However, it is notable that the Chinese subjects showed no effect of plausibility at this position. In their case, the effect was confined to the filler noun (*meal*). This could also reflect a problem with argument substitution at the verb, but with the filler noun being identified as superfluous, rather than the post-verbal noun. The Germans also showed evidence of a similar effect (but again, confined to a minority of subjects), but only for the Chinese was the effect of plausibility confined entirely to the filler noun. It should also be noted that in Experiment 1, only the Chinese group failed to show an effect of plausibility on on-line stop making sense decisions at and following the post-verbal noun, whilst exhibiting the same kind of slow-down in reading times in this region as the other groups. It appears, therefore, that whilst the Chinese do experience processing difficulty when the post-verbal noun is encountered, they are unlikely to regard that noun as the direct cause of the processing problem.

One reason for the different behaviour of the Chinese subjects may be that they have a strong preference to treat any post-verbal noun as the direct object. Although objects frequently appear before the verb in Chinese, when they do so they are explicitly marked by the object marker marker *ba* (provided the object is affected by the verb, as for *hit* but not *see*). When processing English, Chinese subjects may transfer this strong preference for post-verbal noun phrases to be interpreted as objects. This would make them unlikely to regard this constituent as problematic, or superfluous, but instead they would tend to have problems with the filler because it remains unattached. In contrast, object positioning is more variable in Korean and German, and hence it may be more likely that a post-verbal noun will remain unattached. This account is of course highly speculative. The essential point is that whether plausibility effects appeared at the filler or the post-verbal noun, it is possible to attribute them to greater thematic processing difficulties at the verb in the Plausible-at-V condition.

It is of course possible that subjects arrived at a coherent and correct interpretation of the sentences, and that their responses reflected a genuine judgement of implausibility or inappropriateness. For example, the post-verbal nouns in the Plausible-at-V condition may simply have been relatively implausible objects of the accompanying verbs (e.g. subjects may have thought that washing a shirt in a bucket, or buying a radio for a car is implausible). However, such differences in plausibility of the correct interpretation should also have been reflected in the native responses, at least to some extent, but there was no sign of any such effect. It will therefore be assumed that implausibility judgements reflect problems with thematic analysis. It should also be noted that overall response rates to these items were very low, and the errors that were occurring were rather sporadic. At the same time, the effects were consistent over items, suggesting that what underlay the responses was a performance problem, rather than item-specific implausibilities.

GENERAL DISCUSSION

Using an on-line word-by-word reading methodology, Experiment 1 demonstrated remarkable similarities between native and non-native processing of sentences involving extraction from adjuncts. All groups showed evidence of being garden-pathed by the potential gap after the first verb, which was taken as evidence for use of a filler-driven strategy for locating potential gap sites. Reading time on the post-verbal noun was reduced when the filler was implausible as the direct object of the verb, suggesting that the plausibility of the filler-gap dependency was computed almost immediately during on-line processing. Given that in the non-native groups the effect of plausibility was almost entirely confined to the post-verbal noun it was argued that plausibility was affecting a thematic, rather than syntactic, reanalysis process. These results therefore showed that even adult learners of English, whose first language forms questions in a radically different manner to English, adopt native-like strategies for identifying gap sites, and also immediately interpret the meaning of potential filler-gap relations. That is, not only is syntactic processing immediate and incremental, but so too is semantic and thematic processing.

Whereas the task used in Experiment 1 was primarily sensitive to whether readers are led down a garden-path during on-line processing, the off-line task used in Experiment 2 was sensitive to their ability to recover from misanalysis. It was here that differences between natives and non-natives emerged. In natives, stop making sense judgements to the critical items were all but eliminated under these conditions, whereas for the non-natives, the tendency to make stop making sense decisions persisted, especially when the filler was plausible as direct object of the verb. Thus, whilst natives and non-natives do not differ with regard to whether they get led down the type of garden-path investigated here, they do appear to differ in their ability to recover from misanalysis, especially when their first interpretation is highly plausible. The reading times from Experiment 1 also provided some evidence that the natives are able to initiate reanalysis earlier than the non-natives, using a combination of the implausibility of the filler as direct object and the presence of the following determiner.

To what extent can we generalise from the present findings to reading in more natural situations? In the case of natives, the pattern of results in Experiment 1 corroborates the results of eye-movement tracking experiments by Traxler & Pickering (1996) and Pickering & Traxler (1998) where the subjects' task was to answer yes/no comprehension questions. This implies that the operation of the filler-driven strategy, and the on-line use of plausibility constraints, are

processes that are automatic enough in native speakers to influence reading behaviour in a range of situations. Experiment 1 showed that similar effects could be obtained for non-natives in a task situation which encourages incremental syntactic and semantic processing, but it does not allow the inference that the relevant processes are sufficiently automatic to produce the same effects in all reading situations. Experiment 2 employed a more natural reading task in that sentences were not presented incrementally. The effect of plausibility on stop making sense judgements suggests that they were utilizing a filler-driven strategy and plausibility constraints in the same way as in Experiment 1. Therefore, the operation of these processes does not appear to be contingent on a word-by-word presentation. What is still not clear is whether they are dependent on a stop making sense task.

In the case of Chinese subjects we have additional evidence that at least the operation of the filler-driven strategy is not dependent on the stop making sense task. In a separate experiment (reported in Williams & Mobius, 1997) the Chinese subjects from Experiment 1, plus an additional 12 subjects of similar characteristics, were required to read sentences one word at a time and answer a comprehension question after each sentence. We compared reading patterns for the following sentences:

(9) Bill wonders [what]$_i$ the manager wants [*t_i] the assistant to put [t_i] in the sales next week.

(10) Bill wonders if the manager wants the assistant to put the suits in the sales next week.

The filler-driven strategy predicts that in sentence (9), readers should slow down on *the assistant* because they initially interpret *what* as the direct object of *wants*. Sentence (10) is the control sentence where the sequence *the manager wants the assistant* does not contain a gap. There were 5 items per condition mixed in with 23 unrelated filler sentences. The Chinese subjects did indeed slow down on the word *assistant* relative to the control condition (the interaction between condition and position was significant in the verb-determiner-noun region, $p<0.05$). This experiment therefore showed that Chinese subjects utilized the filler-driven strategy even when the task did not involve stop making sense judgements. In conjunction with Experiment 2 this suggests that this strategy is automatic, at least for advanced non-natives. What remains to be seen, however, is whether plausibility is utilized even when the task does not involve explicit plausibility judgements, and when the materials do not contain implausible sentences. This must remain an issue for future research. What the present experiments do show is that non-natives can in principle utilize plausibility during on-line processing in a native-like way.

Somewhat contrary to expectations, the present experiments suggested that natives and non-natives differ more in terms of the processes involved in recovery from garden-paths rather than the processes that lead to them. The process of reanalysis has received relatively little attention in the sentence processing literature, with most research being concerned with the structural factors which make some kinds of garden-path structure easier to recover from than others (Fodor & Inoue, 1994; Sturt, Pickering, & Crocker, 1999). However, Pickering & Traxler (1998) provide evidence for an effect of plausibility on the cost of reanalysis that is somewhat similar to that demonstrated here. They compared reading of sentences such as (11) and (12). Readers had more difficulty in the ambiguous region *which upset kids* in (12) than in (11), but the disambiguating region *harmed too many people* was harder in (11) than in (12).

(11) The criminal confessed his sins which upset kids harmed too many people.

(12) The criminal confessed his gang which upset kids harmed too many people.

In (11), *sins* is initially interpreted as the object of *confessed*, and this analysis has to be revised when the main verb *harmed* is encountered. In (12), the implausibility of *gang* as object of *confessed* leads to increased processing difficulty, but the alternative analysis is easier to

derive ultimately, leading to faster processing in the disambiguating region. However, comparisons with unambiguous control sentences showed that (12) still caused difficulty in the disambiguating region. On this basis Pickering & Traxler argue that the implausibility of *gang* as direct object of *confessed* does not always trigger structural reanalysis. However, it does ease the reanalysis process in the disambiguating region. When an initial analysis is highly plausible, readers commit strongly to it, and reanalysis is made more difficult. As they point out, such an effect is amenable to explanation in a constraint-based framework (MacDonald, Pearlmutter, & Seidenberg, 1994) where many different types of information, both syntactic and nonsyntactic, flexibly combine during the interpretation process. When a candidate analysis is supported by plausibility it is harder for it to be overturned by purely structural cues.

A similar effect was obtained in the present experiments, although it was explained in terms of readers' commitment to a particular thematic, rather than syntactic, analysis, and the difficulty of displacing a plausible argument from the thematic structure even in the face of unambiguous syntactic information. Differences between natives and non-natives in their ability to recover from a plausible misanalysis are presumably a reflection of differences in the balance between syntactic and non-syntactic sources of information. In the Plausible-at-V condition the structural cues in the adjunct extraction structure (beginning with the determiner after the verb) had to overcome the combined strength of the filler-as-direct-object syntactic analysis, plus semantic plausibility. The data from Experiment 1 showed that this was clearly problematic for all groups, but only for the non-natives in Experiment 2 did the syntactic evidence in favour of the adjunct extraction occasionally completely fail to win out, suggesting a relative weakness in this source of information. In the Implausible-at-V condition, the syntactic cues to the adjunct extraction structure were sufficiently strong to trigger a correct thematic analysis because this analysis was supported by plausibility, the filler being more plausible as an adjunct than a direct object. However, here too there is evidence that the natives gave higher priority to syntactic cues. They showed slower reading times on the determiner in this condition, suggesting that the implausibility of the filler as direct object enabled this syntactic cue to trigger rapid reanalysis. In contrast, the non-natives only showed reading time differences at the noun where the adjunct analysis was supported by additional syntactic information in form of the noun itself, and plausibility, the noun being a more plausible direct object than the filler. Thus, the overall pattern of results can be best interpreted within a general constraint satisfaction approach to sentence processing (MacDonald et al., 1994) along with the assumption that the balance between semantic and syntactic cues differs between natives and non-natives.

Interestingly, though, the lack of difference in the qualitative pattern of results between Germans on the one hand and Chinese and Koreans on the other suggests that the nature of the L1 has little impact. In the Introduction we laid out a number of reasons for expecting divergent reading behaviour in the non-native groups. The absence of traces in wh-questions in Chinese and Korean may have been expected to cause problems for the acquisition of a filler-driven strategy compared to Germans. Chinese might have been expected to be more sensitive to plausibility constraints than either Koreans or Germans because of the reliance on this type of information in their native language. None of these predictions were born out. All non-native groups appeared to use a filler-driven strategy, and all non-native groups showed evidence of over-commitment to a plausible first analysis. Thus, at the level of proficiency of the present subjects, specific influences from the nature of the first language could not be identified. The only evidence for group differences was between natives and non-natives. Whether this reflects a difference in degree of exposure, or some unique characteristic of native language processing, will not be clear until non-natives with even higher levels of exposure to the L2 are investigated.

The hypothesis that wh-questions might cause difficulty for Chinese and Korean learners of English was motivated by the assumption that the filler-driven strategy requires sensitivity to a grammatical symbol, in the form of a trace, that is not utilized in the L1 grammar. Does the finding that these groups utilize a filler-driven strategy mean that they have acquired this grammatical concept? Pickering & Barry (1991) argue that human processing of gap structures can be readily accounted for by models of parsing which do not appeal to traces at all, but which instead simply link arguments to subcategorizers by a process of 'direct association'. Like much work on processing gap structures in English, the present results are compatible with either type of approach to parsing because the gap occurred at a point in the sentence where arguments were assigned to thematic roles. This means that the syntactic process of gap filling can not be distinguished from that of mapping arguments onto thematic roles. In fact, all models of parsing for English make the assumption that the human sentence processor attempts to map arguments onto thematic roles at the earliest opportunity. This is instantiated as the principle of Generalized Theta Attachment in Pritchett (1992), as well as the principle of 'Direct Association' in Pickering & Barry (1991). As mentioned in the Introduction, topic movement and scrambling can result in displaced objects in both Chinese and Korean, and so the general process of relating an argument to the theme role of a subsequent verb is not necessarily so unusual for a speaker of these languages. The fact that they can perform this operation in English does not necessarily mean that they have acquired the notion of a trace. The only way of testing this would be to examine structures where the trace position is not adjacent to the relevant subcategorizer, as in certain German constructions (see Clahsen & Featherston, 1999, for evidence of trace effects of this type in native speakers of German). Nevertheless, the present experiments do show that the compulsion for immediate and incremental interpretation at the thematic level also applies to non-native sentence processing, and that this can be achieved even when the precise syntactic cues which control the mapping from surface form to thematic roles are very different from the first language. Where even quite advanced non-natives may encounter problems is not in following garden-paths, or appreciating their semantic plausibility, but in retreating from them even in the face of unambiguous evidence that they have reached a dead-end.

Revision accepted November 16th 2000

REFERENCES

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. Chichester: Wiley.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The Crosslinguistic Study of Sentence Processing* (pp. 3-73). Cambridge: Cambridge University Press.

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language, 34*, 774-806.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Clahsen, H., & Featherston, S. (1999). Antecedent priming and trace positions: Evidence from German scrambling. *Journal of Psycholinguistic Research, 28*, 415-437.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior, 19*, 450-466.

Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: evidence from eye-movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 555-568.

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry, 9*, 427-474.

Fodor, J. D., & Inoue, A. (1994). The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research, 23*(5), 407-434.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language and Linguistic Theory, 5*, 519-559.

Frazier, L., & Clifton, C. (1989). Identifying gaps in English sentences. *Language and Cognitive Processes, 4*, 93-126.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *Quarterly Journal of Experimental Psychology, 50A*(1), 119-148.

Gibson, E., Hickok, G., & Shütze, C. T. (1994). Processing empty categories: A parallel approach. *Journal of Psycholinguistic Research, 23*(5), 381-405.

Gorrell, P. (1995). *Syntax and Parsing*. Cambridge: Cambridge University Press.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*, 25-38.

Hickok, G., Canseco-Gonzalez, E., Zurif, E., & Grimshaw, J. (1992). Modularity in locating gaps. *Journal of Psycholinguistic Research, 21*, 545-561.

Hoover, M. L., & Dwivedi, V. D. (1998). Syntactic processing in skilled bilinguals. *Language Learning, 48*(1), 1-29.

Huang, C.-T. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry, 15*, 531-574.

Juffs, A. (1998). Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning, 48*(1), 107-147.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing. *Studies in Second Language Acquisition, 17*, 483-516.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning, 46*(2), 283-326.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review, 99*, 122-149.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676-703.

Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society, 30*(4), 287-289.

Osaka, M., Osaka, N., & Groner, R. (1993). Language-independent working memory: Evidence from German and French reading span tests. *Bulletin of the Psychonomic Society, 31*, 117-118.

Pearlmutter, N. J., & MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language, 34*, 521-542.

Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes, 6*, 229-259.

Pickering, M. J. (1994). Processing local and unbounded dependencies: A unified account. *Journal of Psycholinguistic Research, 23*(4), 323-352.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(4), 940-961.

Pritchett, B. L. (1992). Parsing with grammar: Islands, heads, and garden paths. In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition, and processing* (pp. 321-349). Dordrecht: Kluwer Academic Publishers.

Stowe, L. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and Cognitive Processes, 1*, 227-245.

Stowe, L. A., Tanenhaus, M. K., & Carlson, G. (1991). Filling gaps on-line: Use of lexical and semantic information in sentence processing. *Language and Speech, 34*(4), 319-340.

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Sructural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language, 40*, 136-150.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language, 35*, 454-475.

Walter, H. C. (2000). *The involvement of working memory in reading in a foreign language.* Unpublished Doctoral dissertation, University of Cambridge.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic Theory and Psychological Reality* . Cambridge Mass.: MIT Press.

Williams, J. N., & Mobius, P. (1997). Syntactic processing strategies in a second language. *University of Cambridge Working Papers in English and Applied Linguistics, 4*, 173-208.

APPENDIX

Plausible-at-V / Implausible-at-V

Which car did the tourist buy the radio for two months ago?
Which friend did the tourist buy the radio for two months ago?

Which girl did the man push the bike into late last night?
Which river did the man push the bike into late last night?

Which machine did the mechanic fix the wheel with two weeks ago?
Which customer did the mechanic fix the wheel for two weeks ago?

Which parcel did the secretary find the bomb in early this morning?
Which floor did the secretary find the bomb on early this morning?

Which relative did the farmer kill the chicken for two weeks ago?
Which stick did the farmer kill the chicken with two weeks ago?

Which ladder did the man repair the roof with during the holidays?
Which friend did the man repair the roof for during the holidays?

Which friend did the gangster hide the car for late last night?
Which cave did the gangster hide the car in late last night?

Which dog did the farmer chase the sheep with early this morning?
Which hill did the farmer chase the sheep up early this morning?

Which station did the architect build the hotel beside during the summer?
Which mountain did the architect build the hotel on during the summer?

Which meal did the chef cook the meat for during the afternoon?
Which pot did the chef cook the meat in during the afternoon?

Which bucket did the lady wash the shirt in early this morning?
Which soap did the lady wash the shirt with early this morning?

Which patient did the doctor examine the blood for early this morning?
Which lab did the doctor examine the blood in early this morning?

Which baby did the boy drop his toys on just after lunch?
Which hole did the boy drop his toys in just after lunch?

Which lorry did the thief crash his car into late last night?
Which wall did the thief crash his car into late last night?

Which toy did the boy break the window with in the afternoon?
Which stone did the boy break the window with in the afternoon?

Which machine did the woman clean the floors with last night?
Which detergent did the woman clean the floors with last night?